

Conditional Elimination through Code Duplication

Joachim Breitner*

June 20, 2011

We propose an optimizing transformation which reduces program runtime at the expense of program size by eliminating conditional jumps.

1 Preface

1.1 Motivation

In a variety of cases, code is written in a way that in one execution, a conditional execution is evaluated several time. Situations where this may be happening include the following:

- Repeated use of the ternary operator ($\cdot ? \cdot : \cdot$) with a common conditional expression.
- An if-then-else statement inside a loop, where the condition is loop invariant.
- Use of macros or inlined functions provided by a library that include conditional expression.
- Conditional jumps implicitly inserted by the compiler due to short-circuit logic.
- Naive code mechanically generated from another source via tools such as parser generators, or compilers of higher languages that compile to C and then invoke a C compiler.
- Conditionals introduced by earlier compilation passes, such as the Partial Dead Code pass conceived by Bodík and Gupta [BG97] are likely to make other conditionals redundant. In fact, the PDE paper recommends a “branch elimination” step without giving the details of this. CECD can serve as an implementation of this step.

*e-mail: mail@joachim-breitner.de

In some of these cases the programmer might be able to eliminate the redundant conditional expression by himself, but often at the cost of less readable code or repetition, such as two instances of the loop mentioned in the second bullet. In other cases, such as the library-provided macros or the generated code, it is not feasible to expect the source code to be free of redundant conditionals. Therefore it is desirable that an optimizing compiler can perform this transformation.

Furthermore, this transformation not only reduces execution time but can enable further optimizations: If the conditional expression is of the form $v == c$ for a variable v and a constant c , a constant propagation pass can replace v by c in the then-branch, which has been enlarged by our optimization. Also, modern computer architectures, due to long pipelines, perform better if fewer conditional jumps occur in the code.

1.2 Outline

In the next section, we explain when a given region to duplicate is valid and how to perform the conditional elimination. Aiming for a very clear, simple and homogeneous presentation, we describe the algorithm in a very general setting. This will possibly introduce dead code. An implementation would either run a dead code elimination pass afterwards or refine the given algorithm as required. The transformation is demonstrated by example.

Section 3 discusses which properties the region should satisfy for the optimization to actually have a positive effect, and how to avoid useless code duplication.

To decide whether to perform the optimization, we give a simple heuristic that selects a region to be duplicated and decides whether the optimization should be performed, weighting the (runtime) benefits weighted against the (code size) cost in subsection 3.2. We also show that a slight more sophisticated approach, which takes profiling information into account, becomes \mathcal{NP} -hard.

Data flow equations for the properties discussed in the preceding two sections are given in 4.

1.3 Acknowledgements

This paper was written for the group project of the CS614 “Advanced Compiler” course at IIT Bombay under Prof. D. M. Dhamdhere. I have had fruitful discussions with him and my fellow group members, Anup Agarwal, Yogesh Bagul and Anup Naik, who subsequently implemented parts of this using the LLVM compiler suite.

2 Conditional elimination

Let e be an expression, which should occur as the condition for a conditional branch in the control flow graph (CFG) of a program, and let v_1, v_2, \dots be the operands of the expression.

Let D be a region of the control flow graph, i.e. $D \subseteq BB$ where BB is the set of basic blocks in the control flow graph.

The region D is *valid* if and only if no basic block body in D contains an assignment to any of the operands v_1, v_2, \dots of e .

The parameters of the optimization are the conditional expression e and any valid set D . The transformation is performed in three steps, where the first step is generic code duplication which does not yet consider the conditional expression, the second step rewires some edges to make the other copies reachable and the last step removes the redundant conditionals. Each step preserves the meaning of the program.

1. (Code duplication) For every basic block $bb_i \in D$, create three copies¹: the *true copy* bb_i^t , the *false copy* bb_i^f and the *unknown copy* bb_i^u . The edges of the graph are modified as follows:
 - An edge between $bb_i \notin D$ and $bb_j \notin D$ is left unchanged.
 - An edge between $bb_i \in D$ and $bb_j \in D$ is reproduced by the three edges bb_i^t to bb_j^t , bb_i^f to bb_j^f and bb_i^u to bb_j^u .
 - An edge between $bb_i \in D$ and $bb_j \notin D$ is reproduced by the three edges bb_i^t to bb_j , bb_i^f to bb_j to bb_i^u to bb_j .
 - An edge between $bb_i \notin D$ and $bb_j \in D$ is changed to an edge from bb_i to bb_j^u .
2. (Conditional evaluation) For every conditional edge from bb_i to bb_j depending on e being true (false) at the end of bb_i , where bb_j is a copy of a node in D , replace it by an edge bb_i to bb_j^t (bb_j^f).
3. (Conditional elimination) For every basic block $bb_i \in D$ which has a conditional branch depending on e being true (false), remove the condition in bb_i^t (bb_i^f), unconditionally follow the true (false) case and remove the other edge.

This algorithm is correct and safe. For correctness, consider an execution path. If the path does not pass any node in D , it is not altered by the above algorithm. If the path passes through D , but only through unknown copies, it is also not altered. If the path eventually reaches a true (false) copy of a node, it must be because of an edge altered in step 2. At that point of execution, the value of e is known to be true (false), and because D is valid, it remains so until the execution path leaves the region D . Any conditional

¹Technically, this is triplication, not duplication.

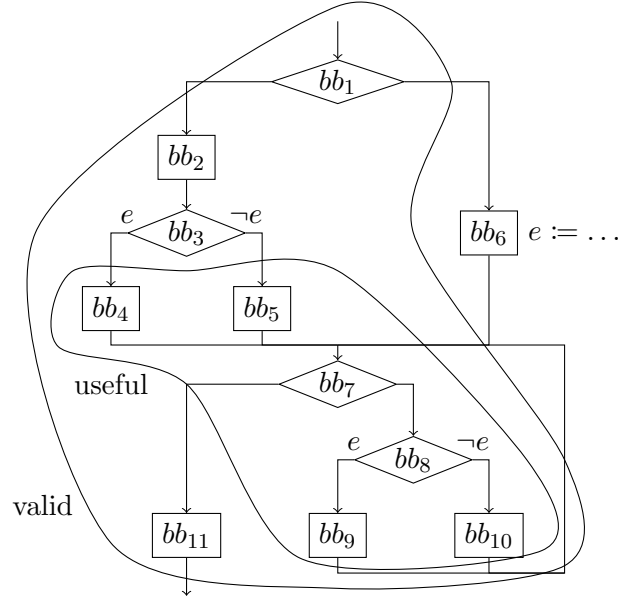


Figure 2: Example control flow graph before CECD

jump skipped because of step 3 is therefore behaving exactly as in the original execution path. .

Safeness follows from the fact that we only copy nodes and remove the evaluation of conditionals, so along no path new instructions are added.

Example

Consider the code fragment in Figure 1 (leaving out any unrelated assignments or expressions). The corresponding control flow graph is given in figure 2. The largest valid region is marked, as well as the largest region if useful nodes. Applying the algorithm with D set to the region of useful nodes, after step 1 we obtain the graph shown in figure 3 on the following page. At this point, the true and false copies are not reachable yet. Steps 2 and 3 modify the edges related to conditional on e , and we reach figure 4 on the next page. This contains a lot of dead code. Removing this in a standard dead code removal pass, we reach the final state 5 on page 6. It can clearly be seen that on every path from entry to exit, the conditional e is evaluated at most once. Also the issue of a while-loop occurrence (in contrast to the optimizer-friendly do-while-loop) is gracefully taken care of.

```

if ... then
| if  $e$  then ... else ...
else
|  $e := \dots$ 
end
while ... do
| if  $e$  then ... else ...
end
...

```

Figure 1: Example code

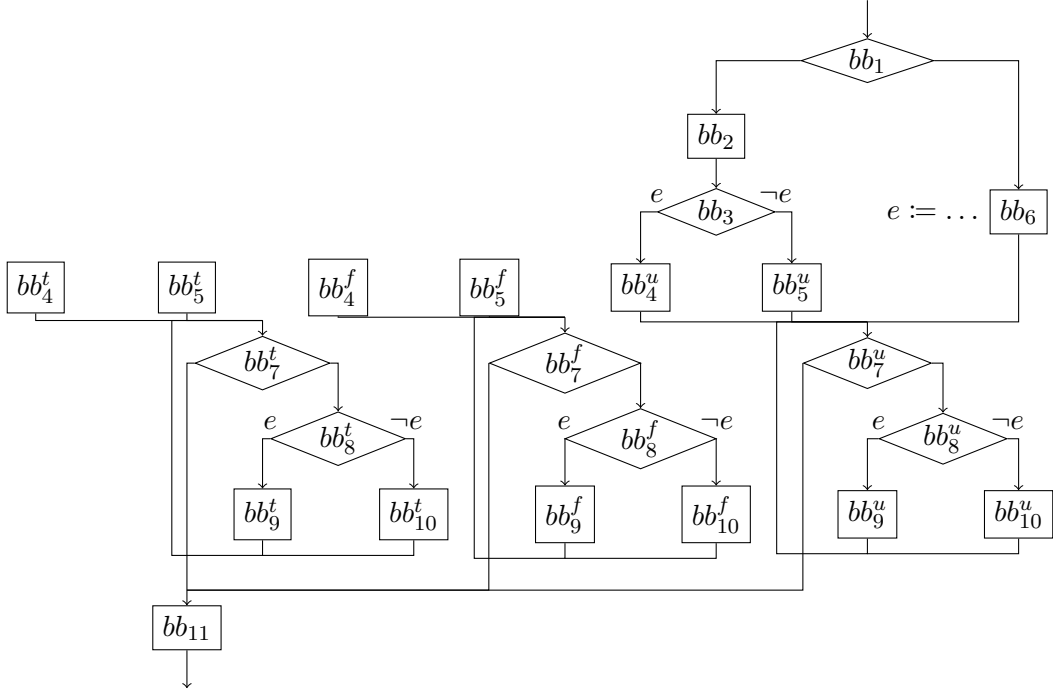


Figure 3: Example control flow graph after code duplication of useful nodes

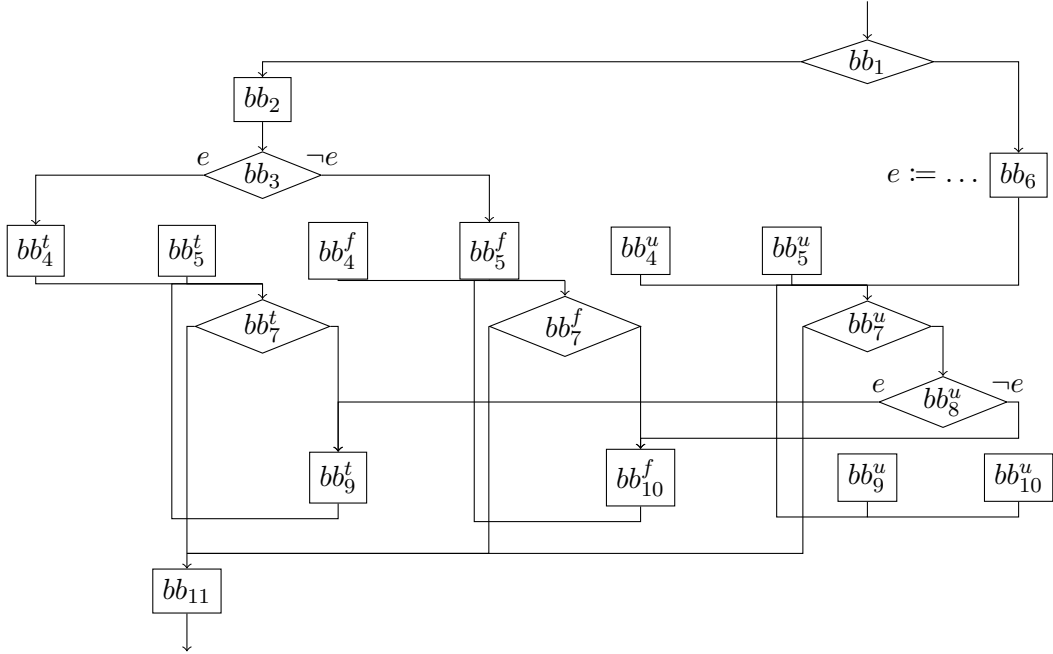


Figure 4: Example control flow graph after conditional evaluation and elimination

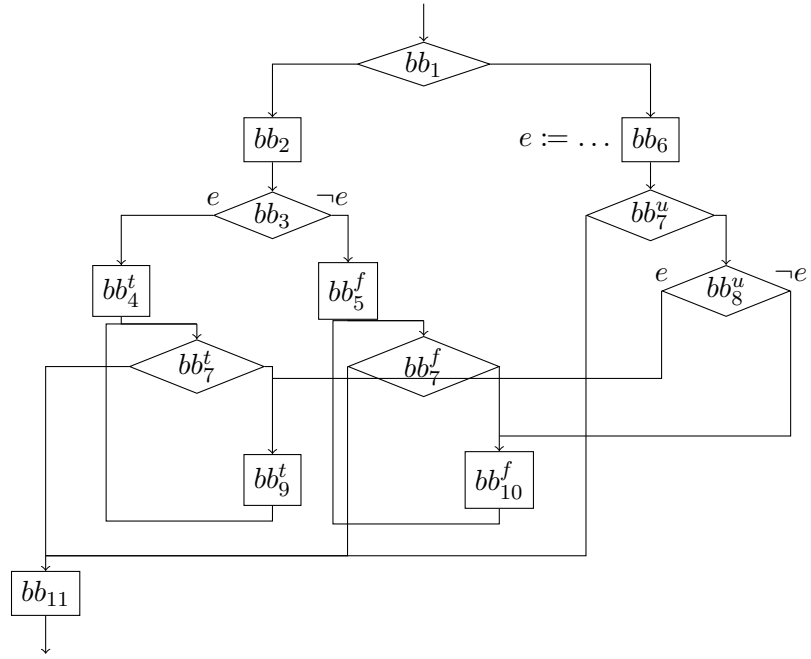


Figure 5: Example control flow graph after conditional evaluation and elimination and dead code elimination

3 The region of duplication

The above algorithm works for any valid region, and validity is a simple local property that is easily checked. But not all valid regions are useful. For example, entry nodes bb_i of the region where no incoming edge depends on e would be duplicated, but only bb_i^u would be reachable. Similarly, exit nodes of the region that do not have a conditional evaluation of e would be copied for no gain.

3.1 Usefulness

Therefore, we can define that a node bb_i in a valid region D to be *useless* if

- on all paths leading to bb_i , there is no conditional evaluation of e followed only by nodes in D or
- no path originating from bb_i reaches an conditional evaluation of e before it leaves the region D .

A node $bb_i \in D$ that is not useless is *useful*.

Uselessness is, in contrast to validity, not a property of the basic block alone but defined with respect to the chosen region D . A basic block may be useless in D but not so in a different region D' . But the property is monotonous: If $D' \subseteq D$ and D is useful in D' , then it is also useful in D .

3.2 Evaluation of a region

For a given conditional expression, there are many possible regions of duplication, and even if we only consider fully useful regions, their number might be exponential in the size of the graph. Therefore we need an heuristic that selects a sensible region or decides that no region is good enough to perform CECD. We split this decision into two independent steps: Region Selection, where the the best region for a particular conditional, for some meaning of “best” is chosen, and Region Evaluation, where it is decided whether CECD should be performed for the selected region.

These decisions have to depend on the intended use of the code. Code for an embedded system might have very tight size requirements and large regions of duplication would be unsuitable, whereas code written for massive numerical calculations may be allowed to grow quite a bit if it removes instructions from the inner loops.

At this point, we suggest a very simple heuristic for Region Selection: To cover as many executions paths as possible, we just pick the largest valid region consisting of useful nodes. The heuristic for Region Evaluation expects one parameter k , which is the number of additional expressions that the program is allowed to grow for one conditional to be removed. Together, this amounts to the following steps being taken:

1. Let D be the largest valid region consisting only of useful nodes.
2. Let R^t , R^f resp. R^u the set of those basic blocks in D , whose true, false resp. unknown copy will be reachable after CECD.
3. Let n be the number of basic blocks in D that contain a conditional evaluation of e , i.e. the number of redundant conditionals.
4. If

$$\sum_{bb_i \in R^t} S(bb_i) + \sum_{bb_i \in R^f} S(bb_i) + \sum_{bb_i \in R^u} S(bb_i) - \sum_{bb_i \in D} S(bb_i) \leq n \cdot k,$$

where k is a user-defined parameter and $S(bb_i)$ is the number of instructions in the basic block bb_i , perform CECD on D , otherwise do not perform CECD for this conditional expression.

A number of improvements to this scheme come to mind:

- The selection heuristic should consider subsets of the largest valid and useful regions as well.
- It should give different weights to conditionals that are completely removed and conditionals that are only partially removed.
- Removal of conditionals in inner loops should allow for a larger increase of code size.
- Given sufficiently detailed execution traces, a more exact heuristic can be implemented. In the next section we see that this easily leads to a \mathcal{NP} -hard problem.

3.3 \mathcal{NP} -hardness of a profiling based Region Selection heuristic

A straight forward extension of the above Region Selection heuristic that takes profiling data in the form of execution traces into account, would maximize the sum $\sum_{bb_i \in E} f(bb_i)$, where E is the set of basic blocks containing an eliminated conditional and $f(bb_i)$ is the number of paths in the execution traces where the conditional in bb_i would be eliminated due to CECD. For simplicity, we assume that an occurrence of a conditional expression does not contribute to the size $S(bb_i)$ of a basic block.

If we have an algorithm that selects the optimal region, we can solve the 0-1 knapsack problem, which is \mathcal{NP} -complete. The specification of this problem is as follows:

Given n items with weight $w_i \in \mathbb{N}$ and value $v_i \in \mathbb{N}$, $i = 1, \dots, n$ and a bound $W \in \mathbb{N}$, find a selection of items $X \subseteq \{1, \dots, n\}$ that maximizes the sum $\sum_{i \in X} v_i$ under the constraint $\sum_{i \in X} w_i \leq W$.

Given such a problem, we construct a control flow graph and profiling data as follows:

- The entry node is bb_s , which contains a conditional expression e . Both conditional branches point to the node bb_r .
- There is one exit node bb_e with a conditional expression e .
- The node bb_r is the root of a binary tree of basic blocks. The inner nodes contain no instructions but conditional jumps with conditional expressions that are pairwise distinct and distinct from e .
- The tree contains n leaf nodes bb_l^i , $i = 1, \dots, n$. The node bb_l^i contains w_i instructions, i.e. $S(bb_l^i) = w_i$ and the profiling data gives a frequency of v_i for the execution path passing through bb_l^i .
- The parameter k is chosen to be W .

A valid and useful region of duplication D in this CFG corresponds to a subset of $X \in 1, \dots, n$ and, if non-empty, includes bb_e , bb_l^i for $i \in X$ and the nodes connecting bb_r with those leaf nodes. Because bb_s dominates all nodes in D , no unknown copies will be generated, and both true and false copies are reachable. The inner nodes of the binary tree and bb_e only contain conditional expressions and thus do not contribute to the size of the duplicated region. Only one redundant conditional occurs, hence $n = 1$. The number of executions of bb_e where the conditional is eliminated is exactly the number of execution paths that pass through one of the leaf nodes in D . Therefore, the constraint imposed by the Region Evaluation heuristic becomes

$$\begin{aligned}
\sum_{bb_i \in R^t} S(bb_i) + \sum_{bb_i \in R^f} S(bb_i) + \sum_{bb_i \in R^u} S(bb_i) - \sum_{bb_i \in D} S(bb_i) &\leq n \cdot k && \iff \\
\sum_{i \in X} S(bb_l^i) + \sum_{i \in X} S(bb_l^i) + 0 - \sum_{i \in X} S(bb_l^i) &\leq 1 \cdot k && \iff \\
\sum_{i \in X} w_i &\leq W
\end{aligned}$$

and the term to be optimized can be transformed as follows:

$$\sum_{bb_i \in E} f(bb_i) = \sum_{i \in X} f(bb_l^i) = \sum_{i \in X} v_i.$$

This concludes the proof of \mathcal{NP} -hardness of this profiling-based heuristic for CECD.

The assumption that conditional expressions do not contribute to the size of a node is not critical: If they do contribute, then this result can still be obtained by a technical modification: Increase k by one and then scale k and the number of instructions in the nodes bb_l^i by a factor larger than the number of all conditional expressions occurring.

4 Data Flow equations

Three properties of basic blocks have been defined so far: Validness, usefulness and, for the heuristics, which copies of the block will be present after dead code removal. The first one is a purely local property, while the others can be obtained by standard data flow analyses. The defining equations are given in this section. $\text{succ}(i)$ is the set of successor nodes of bb_i in the control flow graph, $\text{pred}(i)$ the set of predecessors. We assume that nodes with a conditional jump have exactly two successors, one for true and one for false.

Local properties:

- Valid_i : Basic block bb_i does not contain an assignment to an operator of e .
- TrueEdge_{ij} : An edge $bb_i \rightarrow bb_j$ exists and depends on e being true.
- FalseEdge_{ij} : An edge $bb_i \rightarrow bb_j$ exists and depends on e being false.
- $\text{Expr}_i = \sum_{j \in \text{succ}(i)} \text{TrueEdge}_{ij} + \text{FalseEdge}_{ij}$: e is a conditional expression in bb_i

Determining the largest valid region D of useful nodes:

- $\text{Live}_i = \text{Valid}_i \cdot \sum_{j \in \text{pred}(i)} \text{Expr}_j + \text{Live}_j$
- $\text{Antic}_i = \text{Valid}_i \cdot (\text{Expr}_i + \sum_{j \in \text{succ}(i)} \text{Antic}_j)$
- $D_i = \text{Live}_i \cdot \text{Antic}_i$

Given a valid region D (which may or may not be obtained using our suggested simple heuristic), determining which copies of the nodes therein are reachable:

- $R_i^u = D_i \cdot \sum_{j \in \text{pred}(i)} \neg \text{Expr}_j \cdot (\neg D_j + R_j^u)$
- $R_i^t = D_i \cdot \sum_{j \in \text{pred}(i)} R_j^t + \text{TrueEdge}_{ji}$
- $R_i^f = D_i \cdot \sum_{j \in \text{pred}(i)} R_j^f + \text{FalseEdge}_{ji}$

All given data flow equations are any-path equations and therefore, the values can be initialized to *false* before solving the equations using a standard iterative round-robin or worklist approach.

5 Future work and conclusions

While the “how” of CECD is fully understood, the question of “where” and “when”, i.e. coming up with good heuristics for the selection of the conditional and region of duplication, needs much further investigation. Also, experiments with real code have yet to be conducted to quantify the benefit and suggest good values for the heuristics’ parameters. Another possible improvement would be to not only consider syntactically equal conditions, but also take algebraic identities into account.

The simplicity of the CECD transformation and the fact that it can easily handle complex control flow indicate that it could be an optimization of general interest.

References

- [BG97] BODÍK, Rastislav ; GUPTA, Rajiv: Partial dead code elimination using slicing transformations. In: *SIGPLAN Not.* 32 (1997), May, S. 159–170.
<http://dx.doi.org/10.1145/258916.258930>. – DOI 10.1145/258916.258930.
– ISSN 0362–1340